NCBI's Entrez System

Alex E. Lash, MD
Cancer Imaging Informatics Meeting
Bethesda, MD

Paris, 1830

Georges
Cuvier
(1769-1832)

Étienne
Geoffroy St. Hilaire
(1772-1844)

1830: "Form vs. Function" Debate

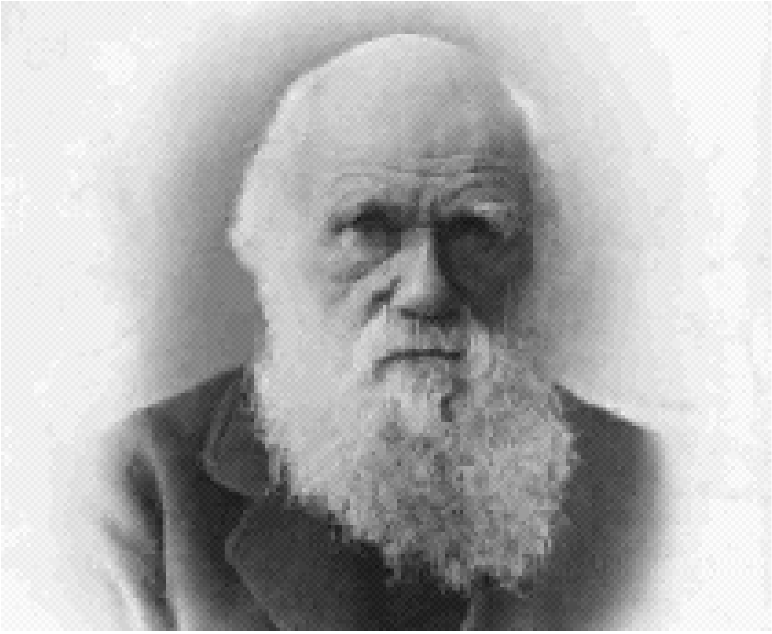
Cuvier

- "form follows function"
- anatomic similarities among vertebrates were due to similar function
- "If there are resemblances between the organs…, it is only insofar as there are resemblances between their functions."

Geoffroy

- "function follows form"
- vertebrates were modifications of a single archetype
- "There is, philosophically speaking, only a single animal."

1859: Darwin on Geoffroy



"Geoffroy St. Hilaire has insisted strongly on the high importance of relative connexion in homologous organs; the parts may change to almost any extent in form and size, and yet they always remain connected together in the same order."

“Pre-hypothesis” Biological Information Collection

Cuvier & Geoffroy
both got here,
but through different reasoning

Collect → Characterize → Relate

where discovery takes place:
patterns are seen
and hypotheses form

A modern example:

<u>Sequencing</u>:
sequence gene

<u>Annotation</u>:
annotate features
such as coding and
non-coding regions

<u>Cross-comparison</u>:
compare sequence
to every other sequence

Today vs. 1830

Biotechnological developments have increased size, scope and speed of "pre-hypothesis" biological information collection.

Collection: overwhelming amount and variety of records

GenBank contains >19 million sequence records and >20 billion bases and doubled in size in the last 16 months

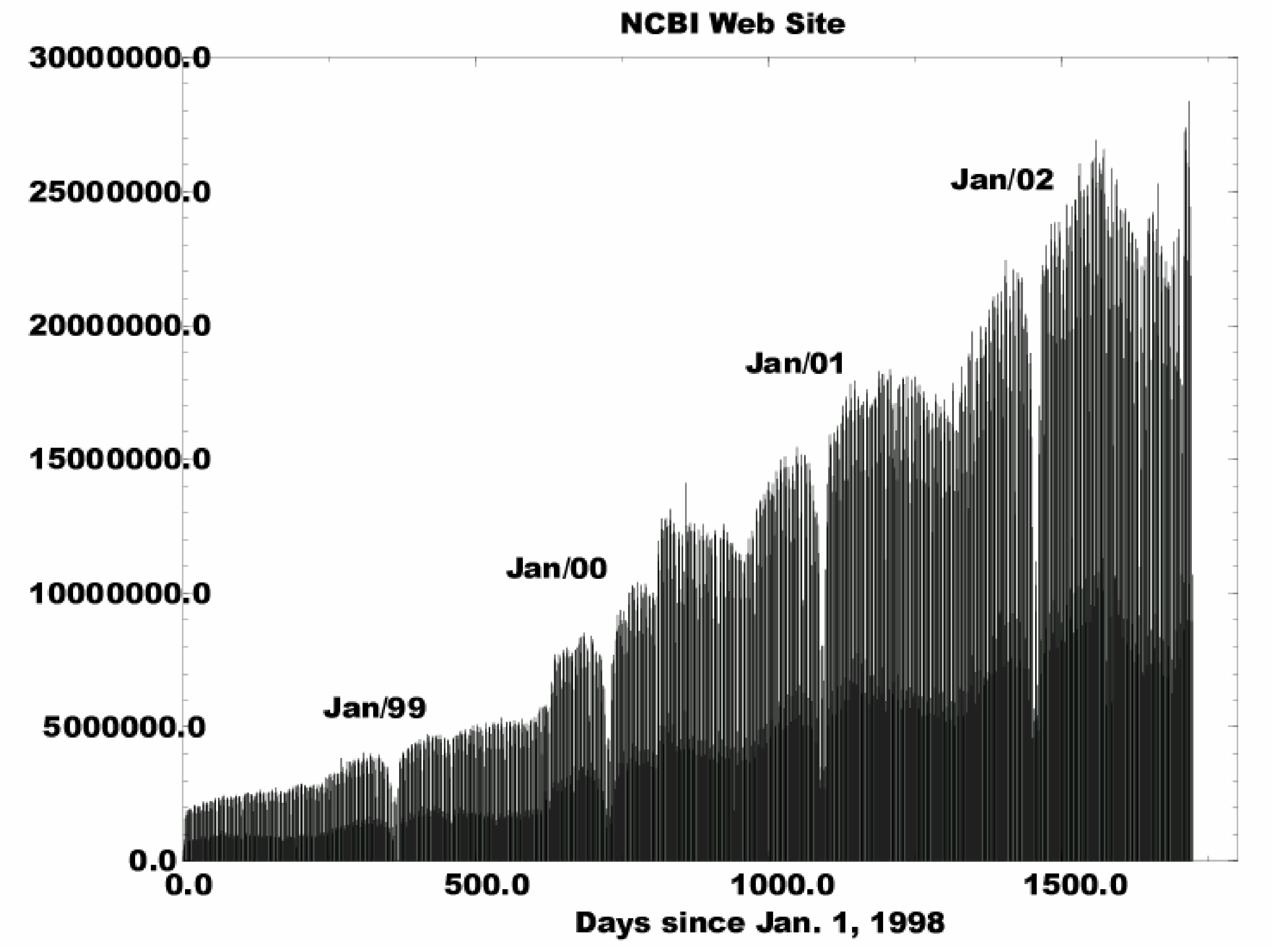
Characterization: increased scope and detail of fields in records

Relation: increased possibility of intra- and inter-database record to record links

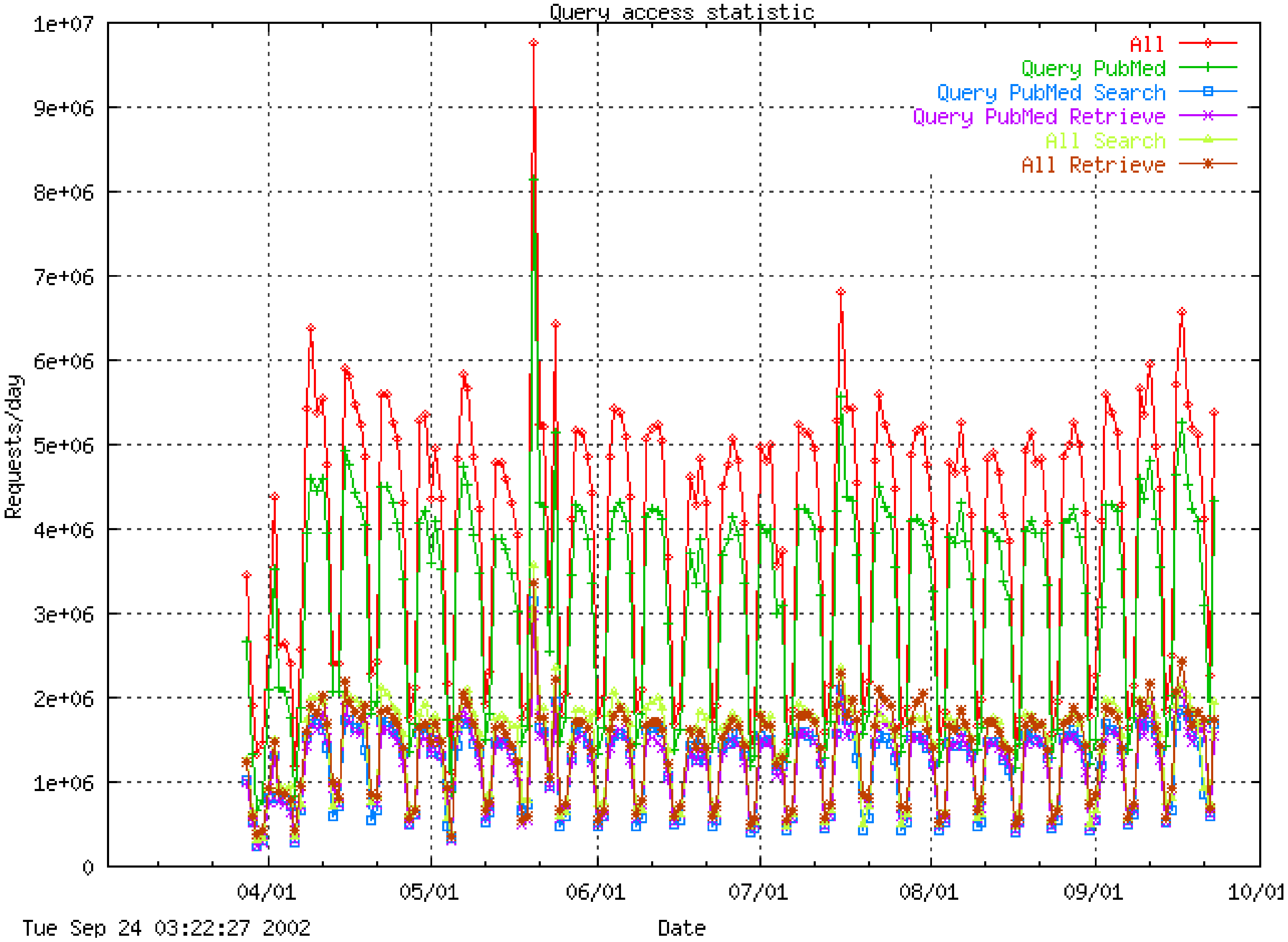
National Center for Biotechnology Information

- Created by Public Law 100-607 in 1988 as part of National Library of Medicine at NIH to:
  - Create automated systems for knowledge about molecular biology, biochemistry, and genetics.
  - Perform research into advanced methods of analyzing and interpreting molecular biology data.
  - Enable biotechnology researchers and medical care personnel to use the systems and methods developed.
- Builders and providers of GenBank, Entrez, Blast, PubMed. Online systems host more than 2 million users per month.
- Center for basic research and training in computational biology.

NCBI Web Hits Per Day



Entrez Hits Per Day



What is Entrez?

Entrez is a scalable and flexible database and interface system constructed and maintained at NCBI.

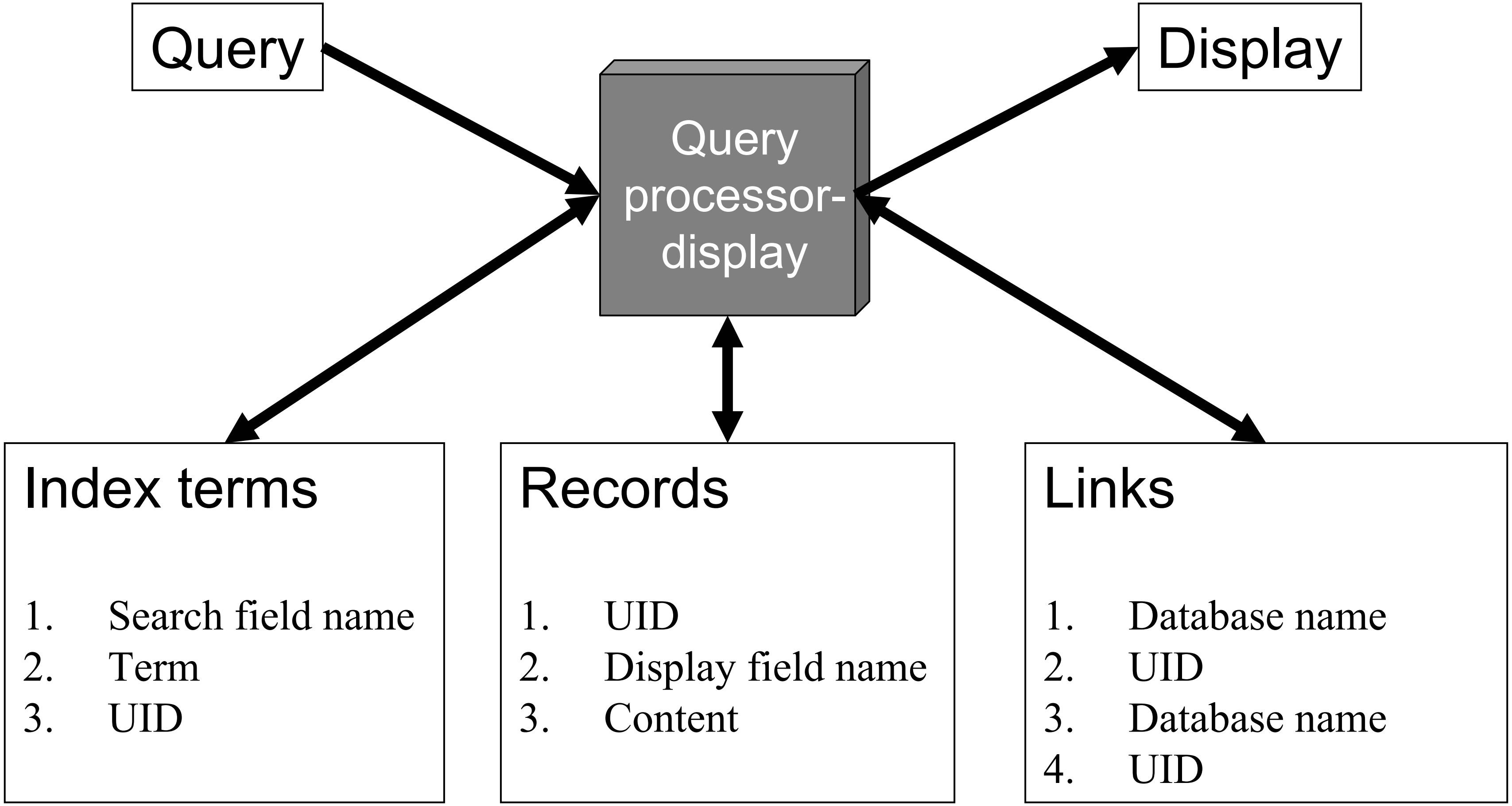
Each Entrez database contains records with pre-specified fields, contains indices on each field, and comes with an interface allowing field-specific, boolean queries.

PubMed is an Entrez database. OMIM is an Entrez database. GenBank nucleotide sequence records are contained in Entrez Nucleotide.

Links can be specified between records within the same Entrez database (intra-database links), or between records in different Entrez databases (inter-database links).

Links can be obvious (eg, identifier matching) or non-obvious (eg, sequence similarity). Non-obvious links generally require examination of the full record and some computation.

Architecture



Query

Query processor-display

Display

Index terms

1. Search field name
2. Term
3. UID

Records

1. UID
2. Display field name
3. Content

Links

1. Database name
2. UID
3. Database name
4. UID

Entrez stats

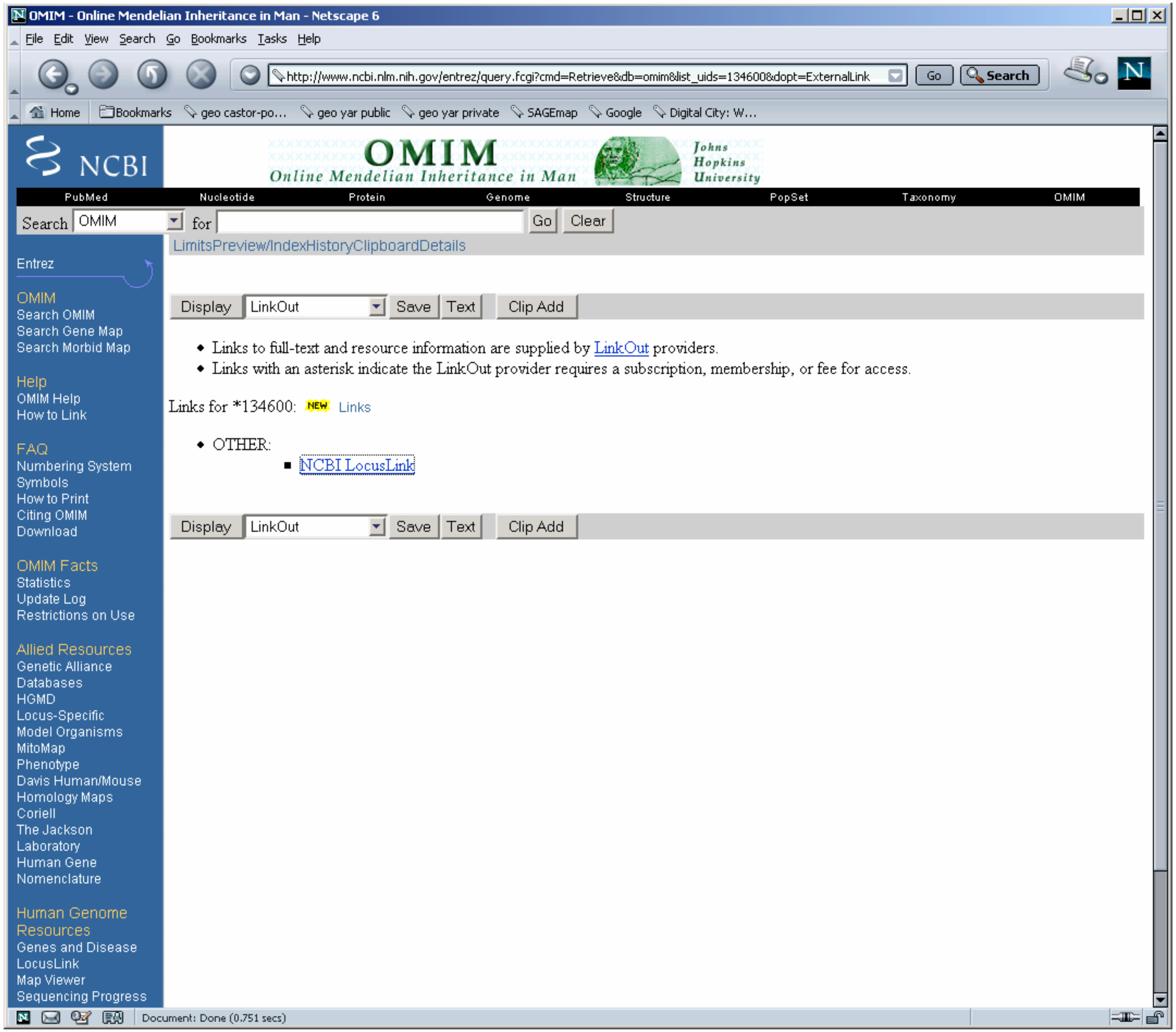
15 Entrez databases

>38 million records

>140 million indexed terms

>6.7 billion intra- and inter-database links

Using Entrez for Discovery - 1



OMIM - Online Mendelian Inheritance in Man - Netscape 6

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=omim&list_uids=134600&dopt=ExternalLink

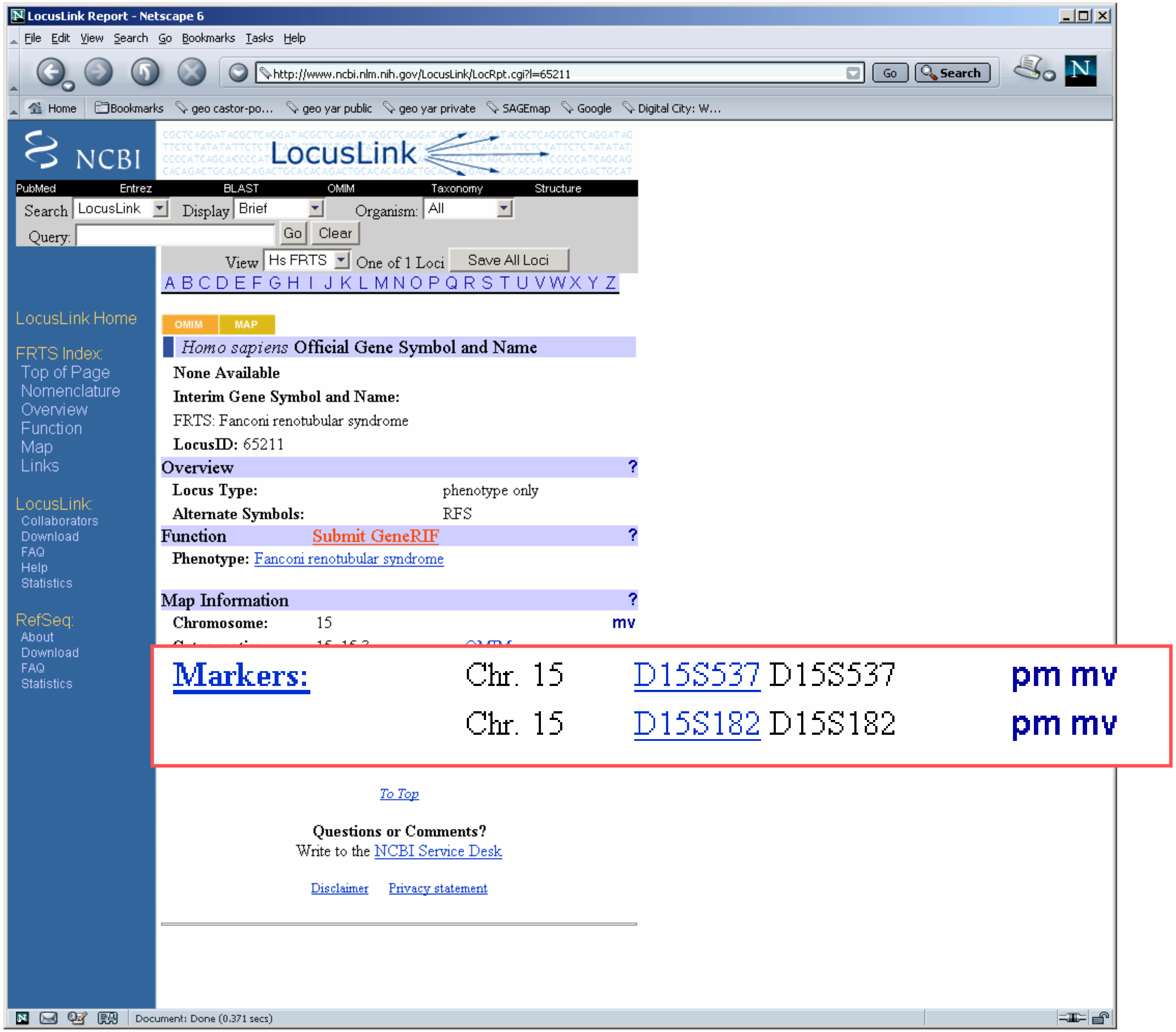
- Links to full-text and resource information are supplied by LinkOut providers.
- Links with an asterisk indicate the LinkOut provider requires a subscription, membership, or fee for access.

Links for *134600: NEW Links

- OTHER:
  - NCBI LocusLink

Document: Done (0.751 secs)

Using Entrez for Discovery - 2



LocusLink Report - Netscape 6
File Edit View Search Go Bookmarks Tasks Help
http://www.ncbi.nlm.nih.gov/LocusLink/LocRpt.cgi?l=65211
Go
Search
Home
Bookmarks
geo castor-po...
geo yar public
geo yar private
SAGEmap
Google
Digital City: W...
NCBI
LocusLink
PubMed Entrez BLAST OMIM Taxonomy Structure
Search LocusLink Display Brief Organism: All
Query:
Go Clear
View Hs FRTS One of 1 Loci Save All Loci
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
LocusLink Home
FRTS Index:
Top of Page
Nomenclature
Overview
Function
Map
Links
LocusLink:
Collaborators
Download
FAQ
Help
Statistics
RefSeq:
About
Download
FAQ
Statistics
OMIM
MAP
Document: Done (0.371 secs)

***Homo sapiens* Official Gene Symbol and Name**

None Available

Interim Gene Symbol and Name:

FRTS: Fanconi renotubular syndrome

LocusID: 65211

Overview ?

Locus Type:	phenotype only
Alternate Symbols:	RFS

Function Submit GeneRIF ?

Phenotype: Fanconi renotubular syndrome

Map Information ?

Chromosome:	15		mv

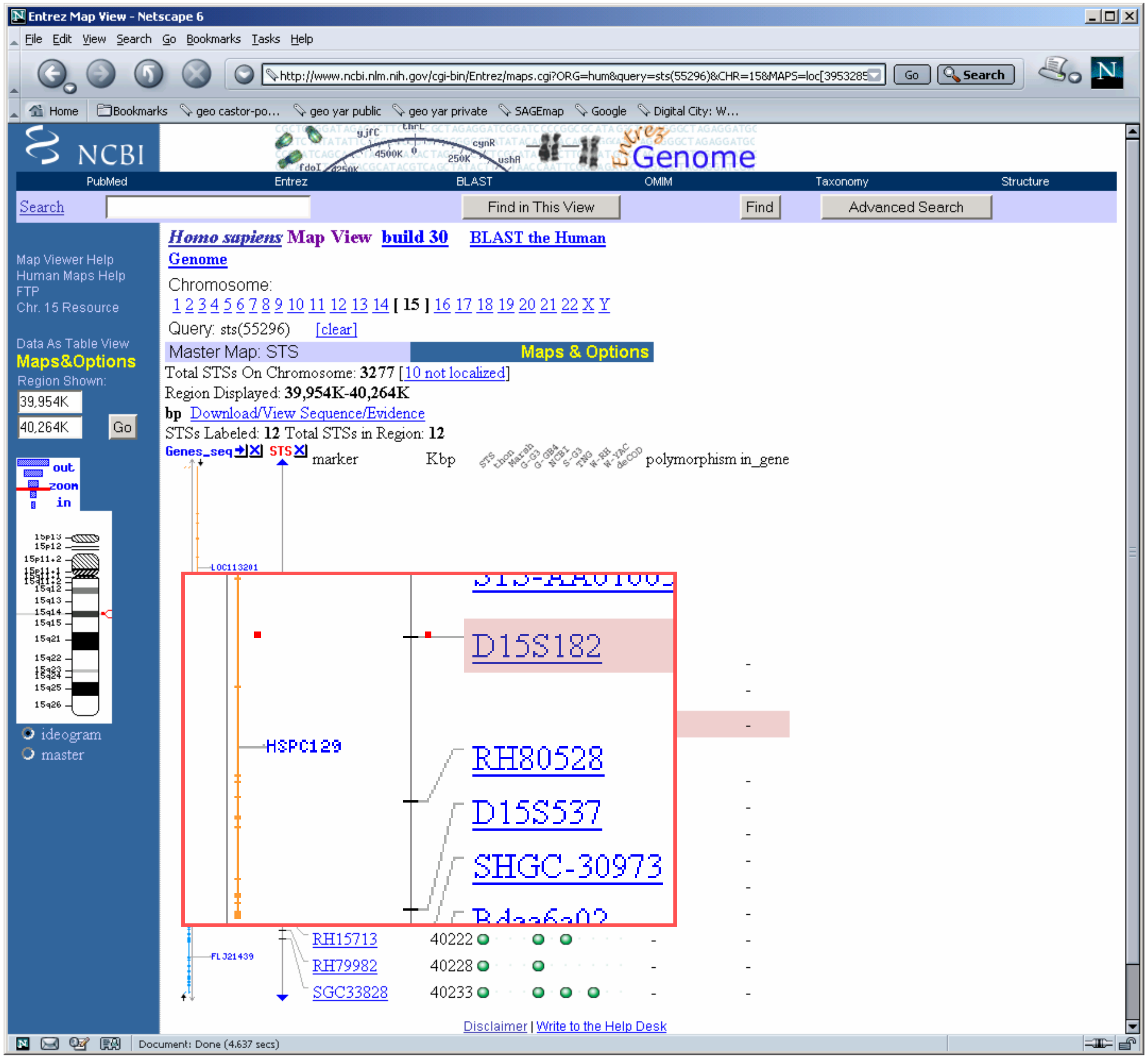
Markers:	Chr. 15	D15S537 D15S537	pm mv
	Chr. 15	D15S182 D15S182	pm mv

To Top

Questions or Comments?
Write to the NCBI Service Desk

Disclaimer Privacy statement

Using Entrez for Discovery - 3



Region Shown:

39,954K

40,264K

Go

Homo sapiens **Map View** **build 30** **BLAST the Human Genome**

Chromosome:

1 2 3 4 5 6 7 8 9 10 11 12 13 14 **[15]** 16 17 18 19 20 21 22 X Y

Query: sts(55296) [clear]

Master Map: STS **Maps & Options**

Total STSs On Chromosome: **3277** [10 not localized]

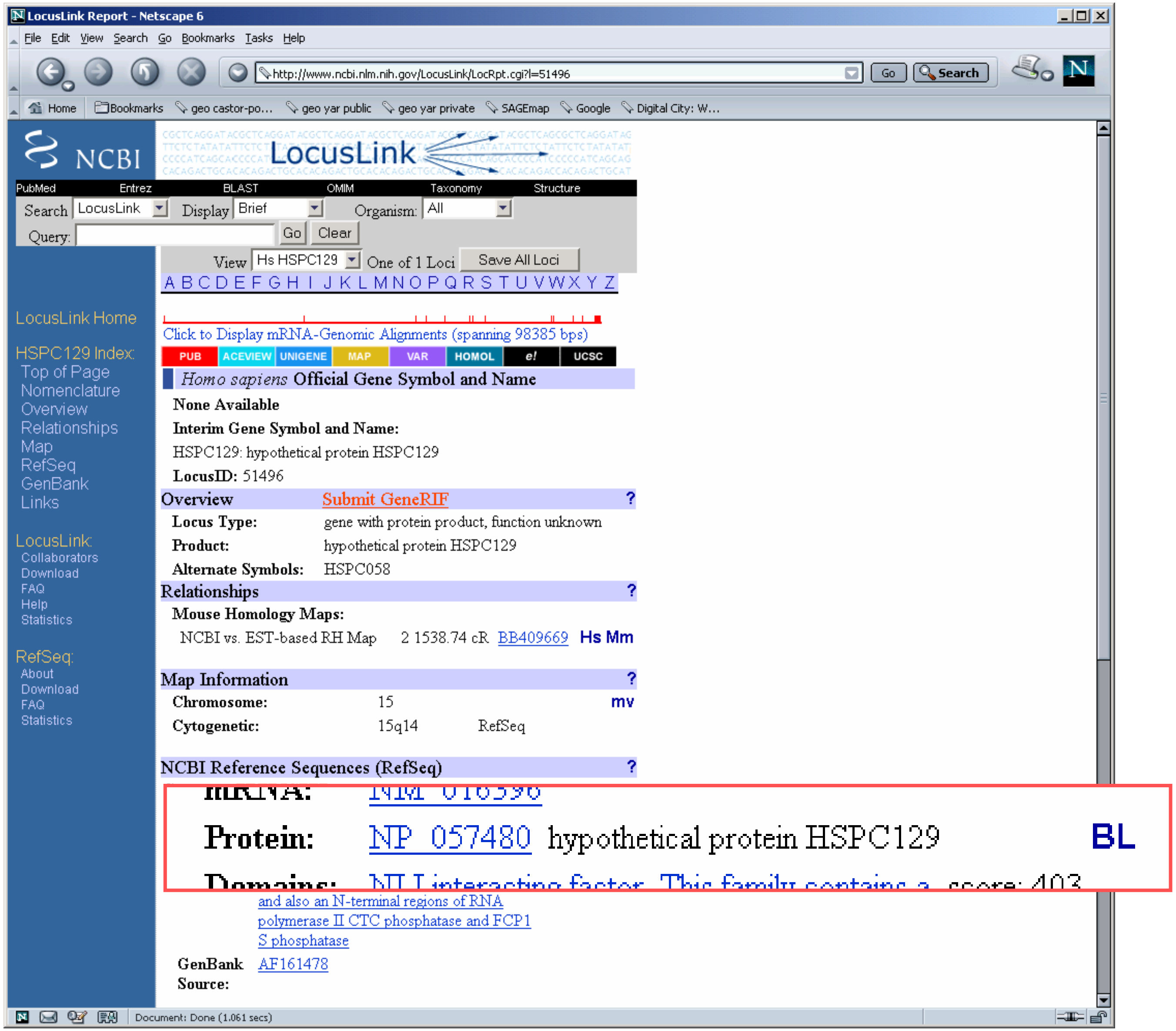
Region Displayed: **39,954K-40,264K**

bp Download/View Sequence/Evidence

STSs Labeled: **12** Total STSs in Region: **12**

marker	Kbp	polymorphism	in_gene
RH15713	40222	-	-
RH79982	40228	-	-
SGC33828	40233	-	-

Using Entrez for Discovery - 4


LocusLink Report - Netscape 6
File Edit View Search Go Bookmarks Tasks Help
http://www.ncbi.nlm.nih.gov/LocusLink/LocRpt.cgi?l=51496
Go
Search
Home
Bookmarks
geo castor-po...
geo yar public
geo yar private
SAGEmap
Google
Digital City: W...
NCBI
LocusLink
PubMed
Entrez
BLAST
OMIM
Taxonomy
Structure
Go
Clear
Save All Loci
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
PUB
ACEVIEW
UNIGENE
MAP
VAR
HOMOL
e!
UCSC
Document: Done (1.061 secs)

Search LocusLink Display Brief Organism: All

Query:

View Hs HSPC129 One of 1 Loci

LocusLink Home

HSPC129 Index:
Top of Page
Nomenclature
Overview
Relationships
Map
RefSeq
GenBank
Links

LocusLink:
Collaborators
Download
FAQ
Help
Statistics

RefSeq:
About
Download
FAQ
Statistics

Click to Display mRNA-Genomic Alignments (spanning 98385 bps)

Homo sapiens Official Gene Symbol and Name

None Available

Interim Gene Symbol and Name:

HSPC129: hypothetical protein HSPC129

LocusID: 51496

Overview Submit GeneRIF ?

Locus Type:	gene with protein product, function unknown
Product:	hypothetical protein HSPC129
Alternate Symbols:	HSPC058

Relationships ?

Mouse Homology Maps:

NCBI vs. EST-based RH Map 2 1538.74 cR BB409669 Hs Mm

Map Information ?

Chromosome:	15		mv
Cytogenetic:	15q14	RefSeq	

NCBI Reference Sequences (RefSeq) ?

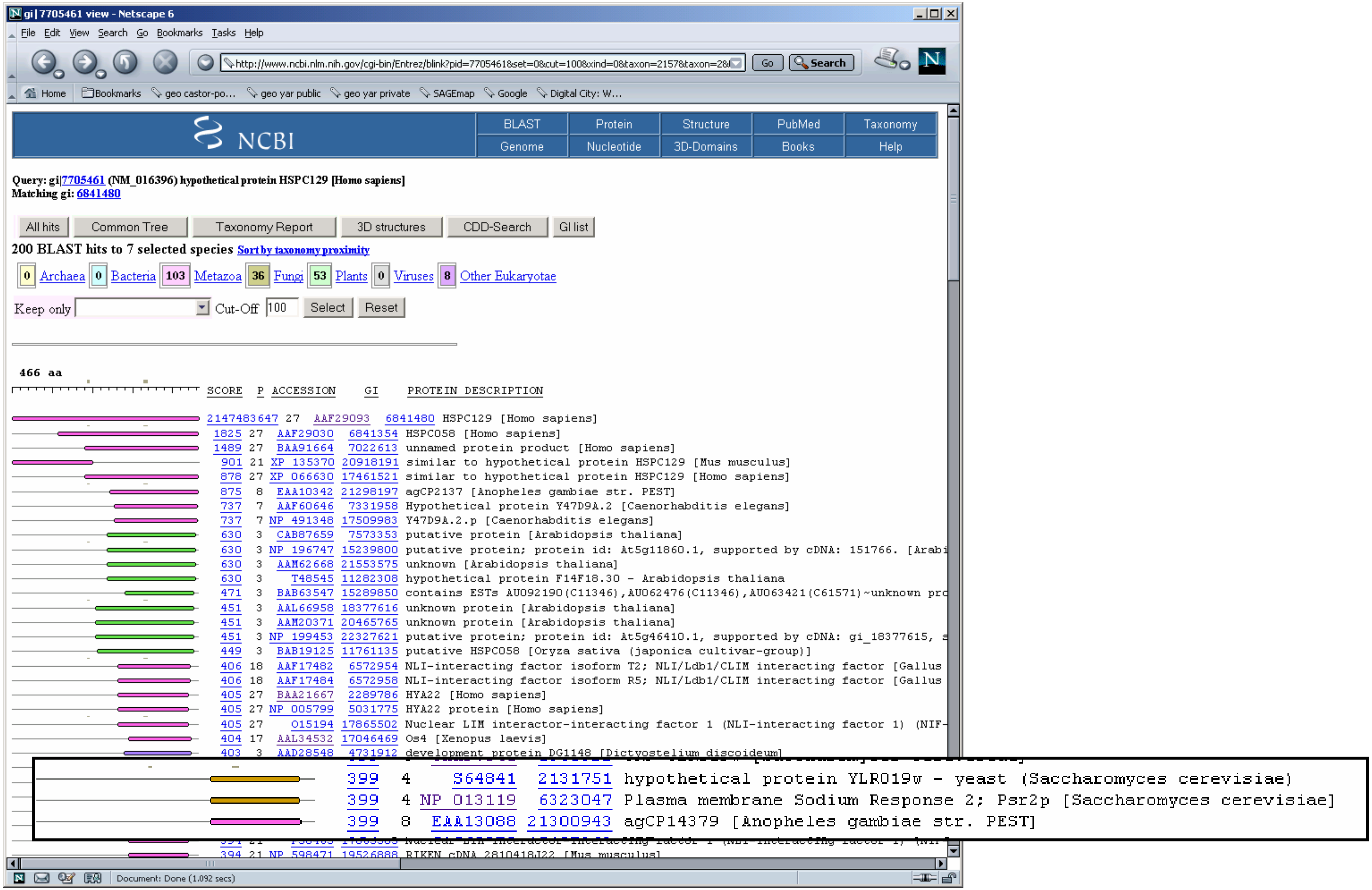
mRNA: NM_016396

Protein: NP_057480 hypothetical protein HSPC129 BL

Domains: NLI interacting factor. This family contains a score: 403 and also an N-terminal regions of RNA polymerase II CTC phosphatase and FCP1 S phosphatase

GenBank Source: AF161478

Using Entrez for Discovery - 5



Query: gi|7705461 (NM_016396) hypothetical protein HSPC129 [Homo sapiens]
Matching gi: 6841480

200 BLAST hits to 7 selected species Sort by taxonomy proximity

0	Archaea	0	Bacteria	103	Metazoa	36	Fungi	53	Plants	0	Viruses	8	Other Eukaryotae

Keep only: □ Cut-Off: 100

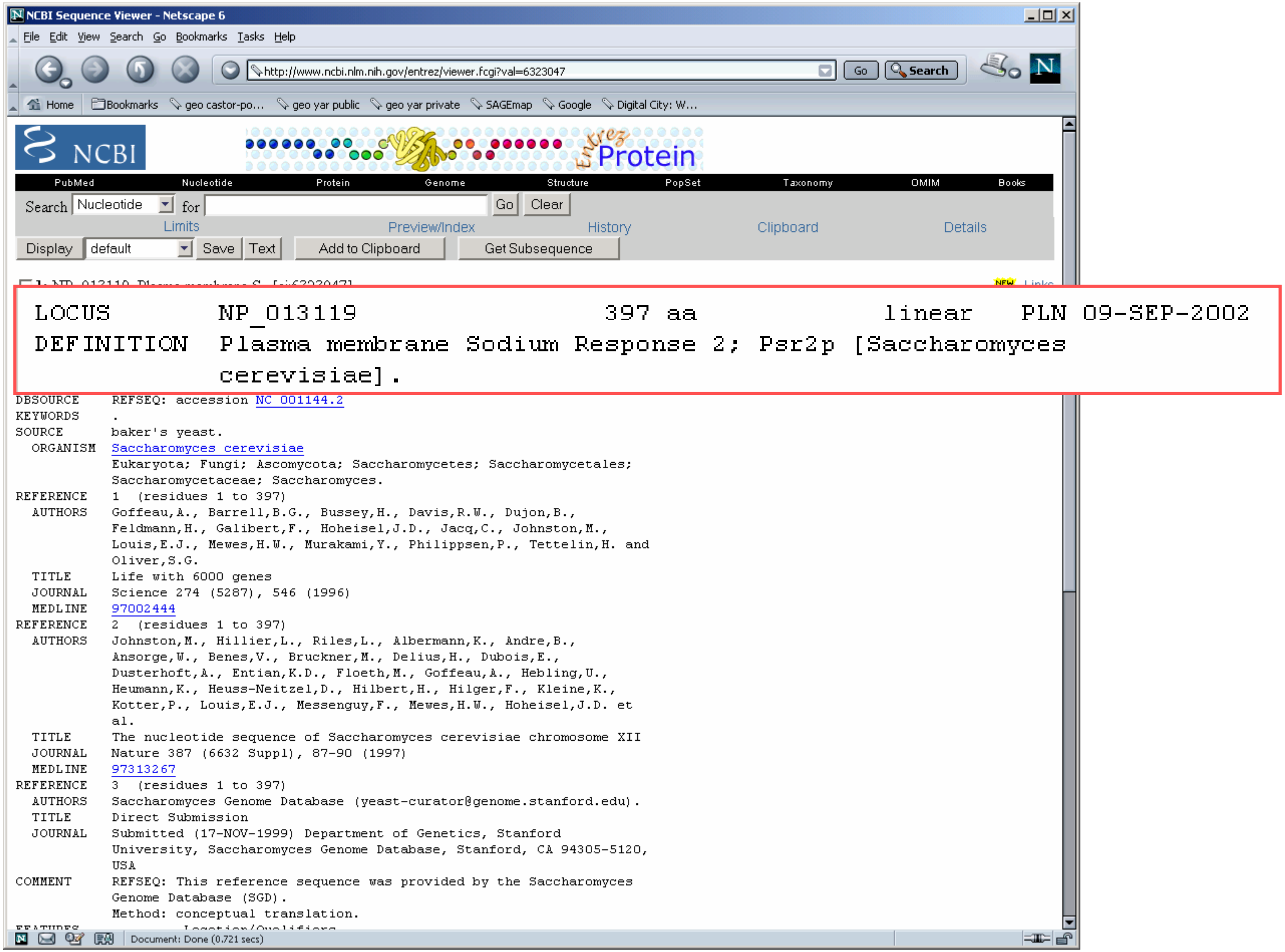
```
466 aa

SCORE       P ACCESSION    GI       PROTEIN DESCRIPTION

2147483647 27  AAF29093   6841480 HSPC129 [Homo sapiens]
      1825 27  AAF29030   6841354 HSPC058 [Homo sapiens]
      1489 27  BAA91664   7022613 unnamed protein product [Homo sapiens]
       901 21 XP 135370  20918191 similar to hypothetical protein HSPC129 [Mus musculus]
       878 27 XP 066630  17461521 similar to hypothetical protein HSPC129 [Homo sapiens]
       875  8  EAA10342  21298197 agCP2137 [Anopheles gambiae str. PEST]
       737  7  AAF60646   7331958 Hypothetical protein Y47D9A.2 [Caenorhabditis elegans]
       737  7 NP 491348  17509983 Y47D9A.2.p [Caenorhabditis elegans]
       630  3  CAB87659   7573353 putative protein [Arabidopsis thaliana]
       630  3 NP 196747  15239800 putative protein; protein id: At5g11860.1, supported by cDNA: 151766. [Arabi
       630  3  AAM62668  21553575 unknown [Arabidopsis thaliana]
       630  3    T48545  11282308 hypothetical protein F14F18.30 - Arabidopsis thaliana
       471  3  BAB63547  15289850 contains ESTs AU092190(C11346),AU062476(C11346),AU063421(C61571)~unknown pro
       451  3  AAL66958  18377616 unknown protein [Arabidopsis thaliana]
       451  3  AAM20371  20465765 unknown protein [Arabidopsis thaliana]
       451  3 NP 199453  22327621 putative protein; protein id: At5g46410.1, supported by cDNA: gi_18377615, s
       449  3  BAB19125  11761135 putative HSPC058 [Oryza sativa (japonica cultivar-group)]
       406 18  AAF17482   6572954 NLI-interacting factor isoform T2; NLI/Ldb1/CLIM interacting factor [Gallus
       406 18  AAF17484   6572958 NLI-interacting factor isoform R5; NLI/Ldb1/CLIM interacting factor [Gallus
       405 27  BAA21667   2289786 HYA22 [Homo sapiens]
       405 27 NP 005799   5031775 HYA22 protein [Homo sapiens]
       405 27    O15194  17865502 Nuclear LIM interactor-interacting factor 1 (NLI-interacting factor 1) (NIF-
       404 17  AAL34532  17046469 Os4 [Xenopus laevis]
       403  3  AAD28548   4731912 development protein DG1148 [Dictyostelium discoideum]
       394 21 NP 598471  19526888 RIKEN cDNA 2810418J22 [Mus musculus]
```

```
399   4     S64841   2131751 hypothetical protein YLR019w - yeast (Saccharomyces cerevisiae)
399   4 NP 013119   6323047 Plasma membrane Sodium Response 2; Psr2p [Saccharomyces cerevisiae]
399   8  EAA13088  21300943 agCP14379 [Anopheles gambiae str. PEST]
```

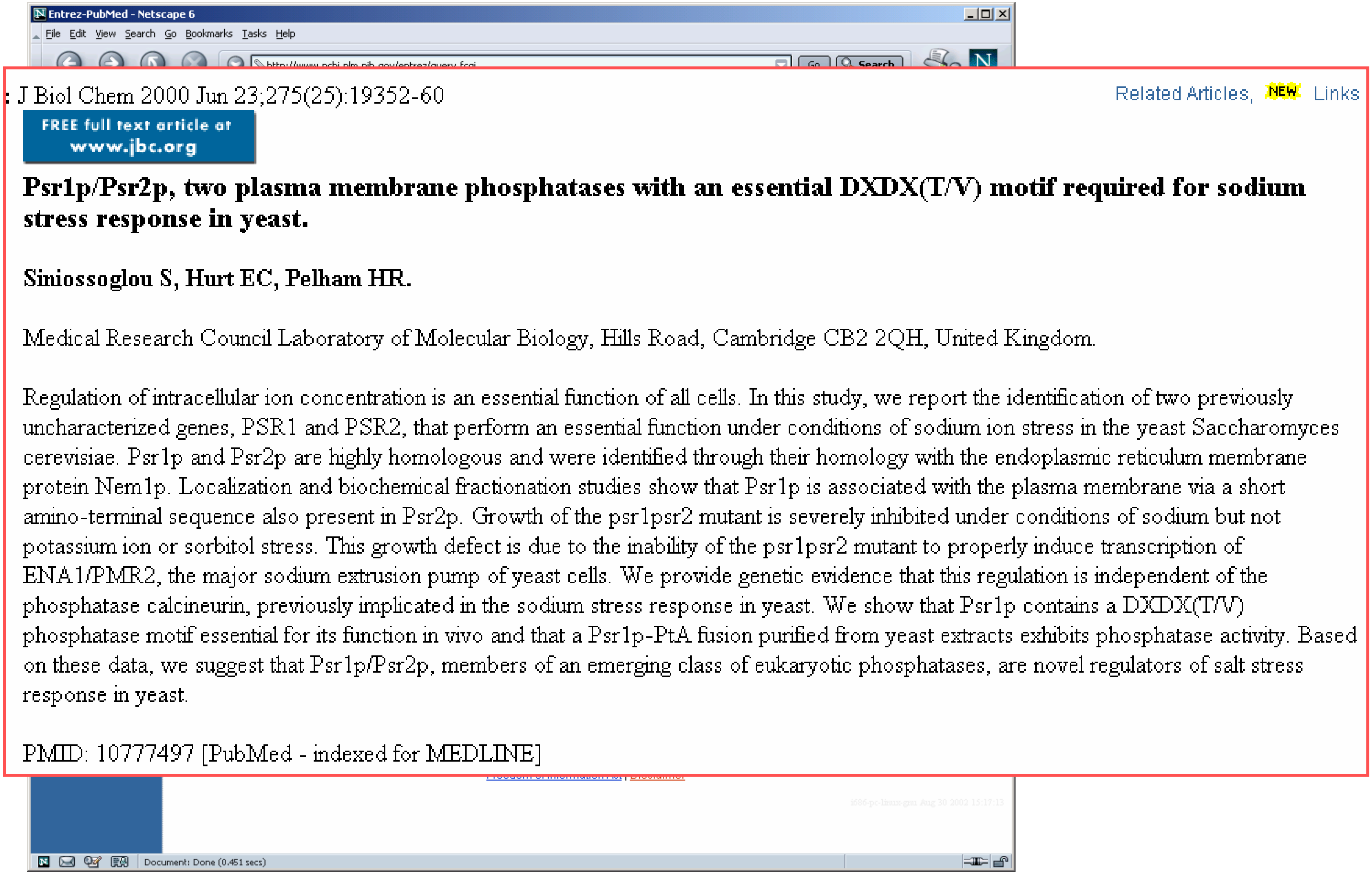
Using Entrez for Discovery - 6



```
1: NP_013119 Plasma membrane S... [gi:6323047]

LOCUS       NP_013119                 397 aa            linear   PLN 09-SEP-2002
DEFINITION  Plasma membrane Sodium Response 2; Psr2p [Saccharomyces
            cerevisiae].
DBSOURCE    REFSEQ: accession NC_001144.2
KEYWORDS    .
SOURCE      baker's yeast.
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycetes; Saccharomycetales;
            Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (residues 1 to 397)
  AUTHORS   Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B.,
            Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M.,
            Louis,E.J., Mewes,H.W., Murakami,Y., Philippsen,P., Tettelin,H. and
            Oliver,S.G.
  TITLE     Life with 6000 genes
  JOURNAL   Science 274 (5287), 546 (1996)
  MEDLINE   97002444
REFERENCE   2  (residues 1 to 397)
  AUTHORS   Johnston,M., Hillier,L., Riles,L., Albermann,K., Andre,B.,
            Ansorge,W., Benes,V., Bruckner,M., Delius,H., Dubois,E.,
            Dusterhoft,A., Entian,K.D., Floeth,M., Goffeau,A., Hebling,U.,
            Heumann,K., Heuss-Neitzel,D., Hilbert,H., Hilger,F., Kleine,K.,
            Kotter,P., Louis,E.J., Messenguy,F., Mewes,H.W., Hoheisel,J.D. et
            al.
  TITLE     The nucleotide sequence of Saccharomyces cerevisiae chromosome XII
  JOURNAL   Nature 387 (6632 Suppl), 87-90 (1997)
  MEDLINE   97313267
REFERENCE   3  (residues 1 to 397)
  AUTHORS   Saccharomyces Genome Database (yeast-curator@genome.stanford.edu).
  TITLE     Direct Submission
  JOURNAL   Submitted (17-NOV-1999) Department of Genetics, Stanford
            University, Saccharomyces Genome Database, Stanford, CA 94305-5120,
            USA
COMMENT     REFSEQ: This reference sequence was provided by the Saccharomyces
            Genome Database (SGD).
            Method: conceptual translation.
FEATURES             Location/Qualifiers
```

Using Entrez for Discovery - 7


Entrez-PubMed - Netscape 6
File Edit View Search Go Bookmarks Tasks Help
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi
Go
Search
Document: Done (0.451 secs)

: J Biol Chem 2000 Jun 23;275(25):19352-60

Related Articles, NEW Links

FREE full text article at www.jbc.org

Psr1p/Psr2p, two plasma membrane phosphatases with an essential DXDX(T/V) motif required for sodium stress response in yeast.

Siniossoglou S, Hurt EC, Pelham HR.

Medical Research Council Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, United Kingdom.

Regulation of intracellular ion concentration is an essential function of all cells. In this study, we report the identification of two previously uncharacterized genes, PSR1 and PSR2, that perform an essential function under conditions of sodium ion stress in the yeast Saccharomyces cerevisiae. Psr1p and Psr2p are highly homologous and were identified through their homology with the endoplasmic reticulum membrane protein Nem1p. Localization and biochemical fractionation studies show that Psr1p is associated with the plasma membrane via a short amino-terminal sequence also present in Psr2p. Growth of the psr1psr2 mutant is severely inhibited under conditions of sodium but not potassium ion or sorbitol stress. This growth defect is due to the inability of the psr1psr2 mutant to properly induce transcription of ENA1/PMR2, the major sodium extrusion pump of yeast cells. We provide genetic evidence that this regulation is independent of the phosphatase calcineurin, previously implicated in the sodium stress response in yeast. We show that Psr1p contains a DXDX(T/V) phosphatase motif essential for its function in vivo and that a Psr1p-PtA fusion purified from yeast extracts exhibits phosphatase activity. Based on these data, we suggest that Psr1p/Psr2p, members of an emerging class of eukaryotic phosphatases, are novel regulators of salt stress response in yeast.

PMID: 10777497 [PubMed - indexed for MEDLINE]

New Entrez Databases

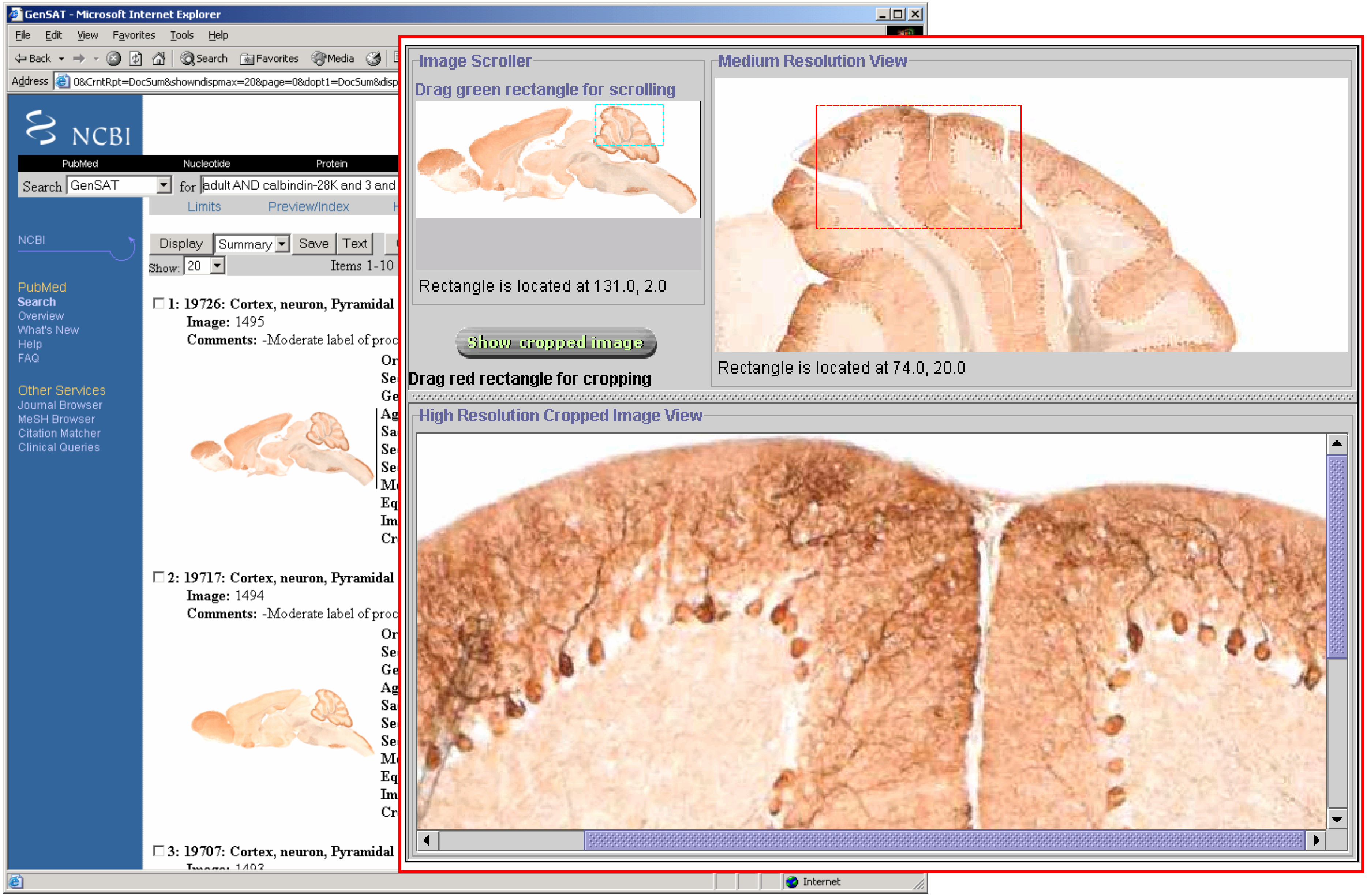
6 new databases in the last year

1. Books: online books
2. GEO: high-throughput gene expression and microarray datasets
3. 3D Domains: structural protein domains from Entrez Structure
4. UniSTS: markers and mapping data
5. CDD: conserved protein domains
6. SNP: single nucleotide polymorphisms

5 new databases on the way

1. UniGene: clusters of sequence similar transcripts
2. Gene: a derivation of LocusLink and Genomes
3. SKY/CGH: spectral karyotyping/comparative genomic hybridization
4. Site Search: search the NCBI web and ftp sites
5. Gensat: *in situ* gene expression in the nervous system of the mouse

Entrez Gensat



Address: 0&CrntRpt=DocSum&showndispmax=20&page=0&dopt1=DocSum&disp

Search: GenSAT for adult AND calbindin-28K and 3 and

Show: 20 Items 1-10

1: 19726: Cortex, neuron, Pyramidal
Image: 1495
Comments: -Moderate label of proc

Or
Se
Ge
Ag
Sa
Se
Se
M
Eq
Im
Cr

2: 19717: Cortex, neuron, Pyramidal
Image: 1494
Comments: -Moderate label of proc

Or
Se
Ge
Ag
Sa
Se
Se
M
Eq
Im
Cr

3: 19707: Cortex, neuron, Pyramidal
Image: 1493

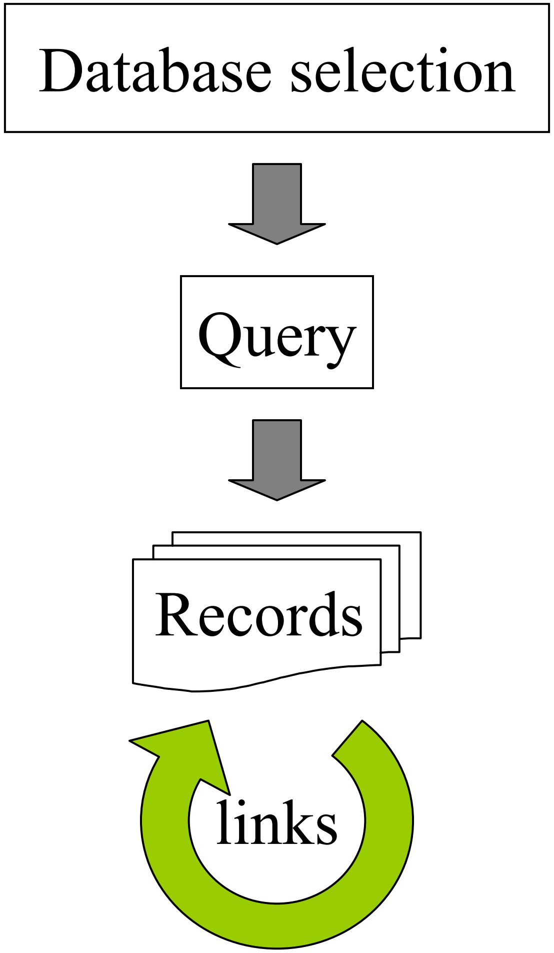
Drag green rectangle for scrolling

Rectangle is located at 131.0, 2.0

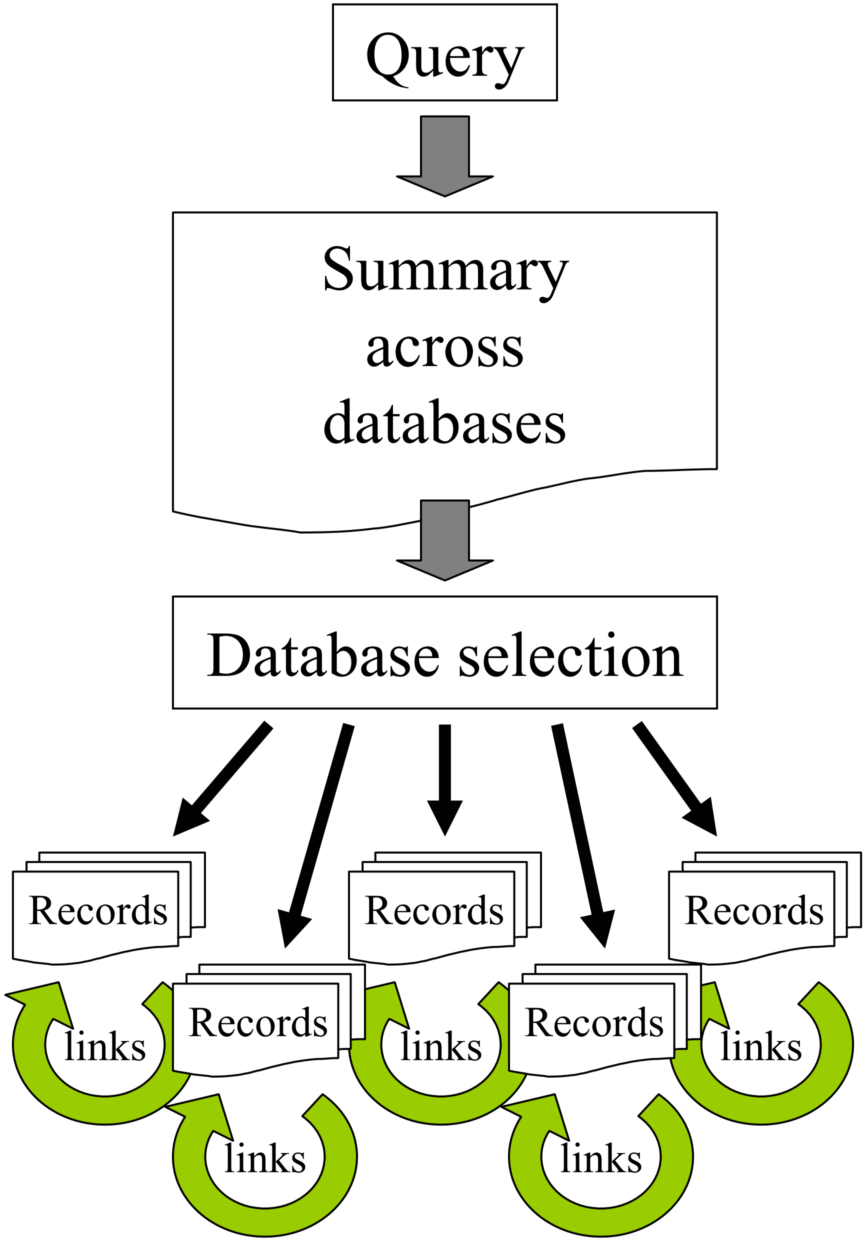
Drag red rectangle for cropping

Rectangle is located at 74.0, 20.0

Current Query Scheme



Global Query Scheme



Query

Summary across databases

Database selection

Records

Records

Records

Records

Records

links

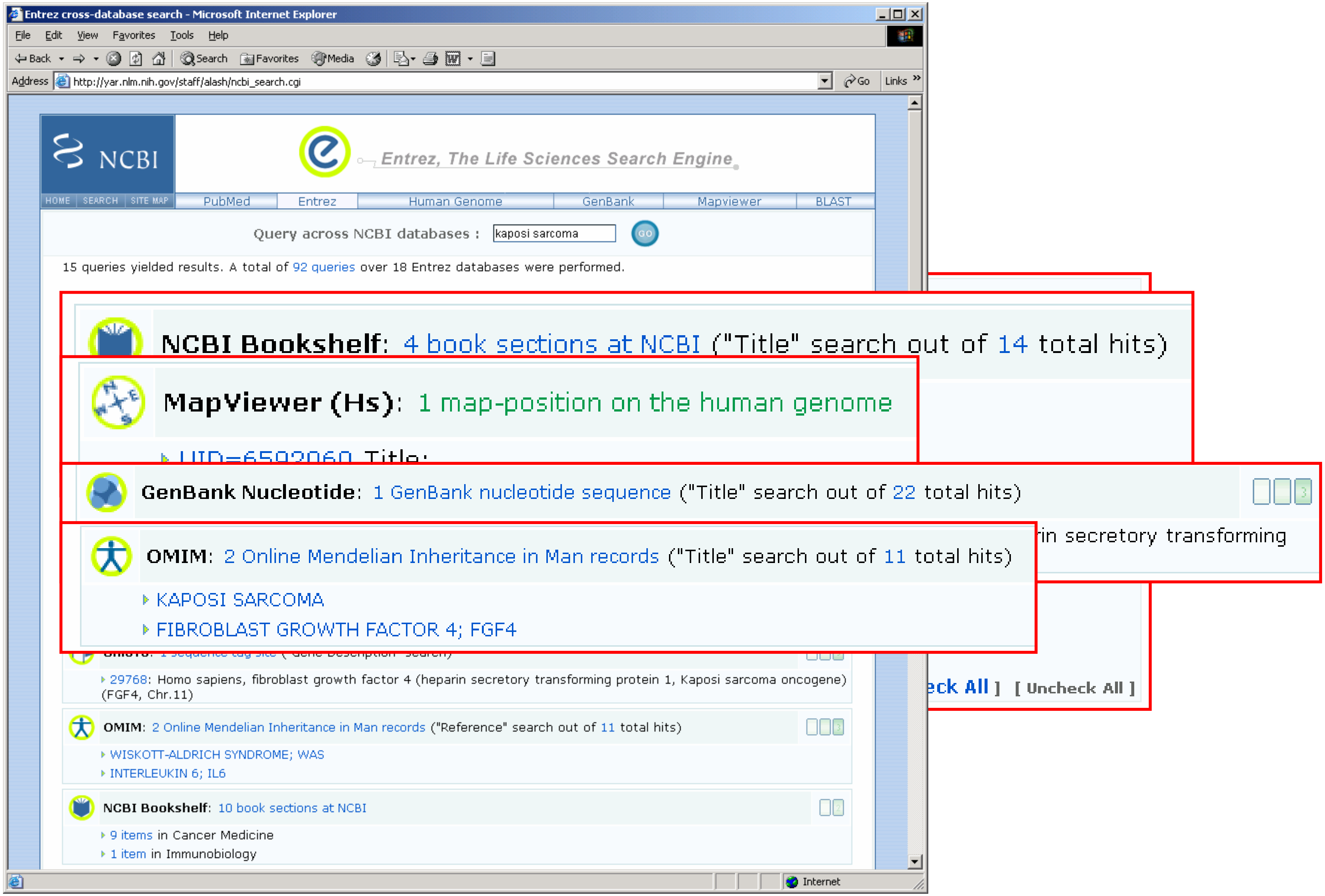
links

links

links

links

Entrez Global Query



Entrez cross-database search - Microsoft Internet Explorer

Address http://yar.nlm.nih.gov/staff/alash/ncbi_search.cgi

Entrez, The Life Sciences Search Engine

Query across NCBI databases : kaposi sarcoma

15 queries yielded results. A total of 92 queries over 18 Entrez databases were performed.

NCBI Bookshelf: 4 book sections at NCBI ("Title" search out of 14 total hits)

MapViewer (Hs): 1 map-position on the human genome

- UID=6592060 Title:

GenBank Nucleotide: 1 GenBank nucleotide sequence ("Title" search out of 22 total hits)

in secretory transforming

OMIM: 2 Online Mendelian Inheritance in Man records ("Title" search out of 11 total hits)

- KAPOSI SARCOMA
- FIBROBLAST GROWTH FACTOR 4; FGF4

eck All] [Uncheck All]

UniSTS: 1 sequence tag site (Gene Description search)

- 29768: Homo sapiens, fibroblast growth factor 4 (heparin secretory transforming protein 1, Kaposi sarcoma oncogene) (FGF4, Chr.11)

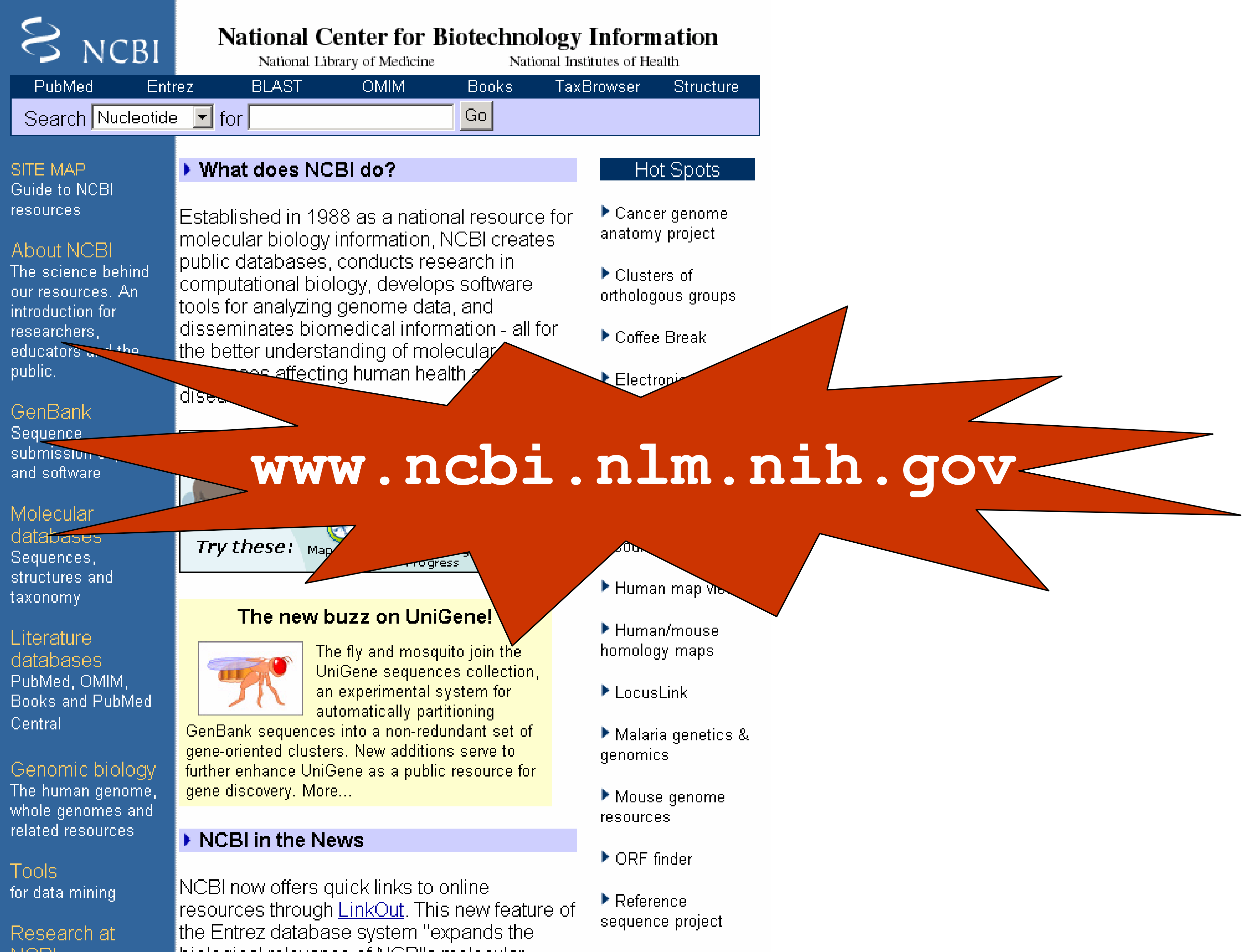
OMIM: 2 Online Mendelian Inheritance in Man records ("Reference" search out of 11 total hits)

- WISKOTT-ALDRICH SYNDROME; WAS
- INTERLEUKIN 6; IL6

NCBI Bookshelf: 10 book sections at NCBI

- 9 items in Cancer Medicine
- 1 item in Immunobiology

File Edit View Favorites Tools Help
Back Search Favorites Media
Go Links
NCBI
HOME SEARCH SITE MAP
PubMed Entrez Human Genome GenBank Mapviewer BLAST
GO
Internet



www.ncbi.nlm.nih.gov

National Center for Biotechnology Information

National Library of Medicine National Institutes of Health

SITE MAP
Guide to NCBI resources

About NCBI
The science behind our resources. An introduction for researchers, educators a... the public.

GenBank
Sequence submission ... and software

Molecular databases
Sequences, structures and taxonomy

Literature databases
PubMed, OMIM, Books and PubMed Central

Genomic biology
The human genome, whole genomes and related resources

Tools
for data mining

Research at NCBI

What does NCBI do?

Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular ... esses affecting human health a... dise...

Try these: Map ... rogress

The new buzz on UniGene!

The fly and mosquito join the UniGene sequences collection, an experimental system for automatically partitioning GenBank sequences into a non-redundant set of gene-oriented clusters. New additions serve to further enhance UniGene as a public resource for gene discovery. More...

NCBI in the News

NCBI now offers quick links to online resources through LinkOut. This new feature of the Entrez database system "expands the biological relevance of NCBI's molecular

Hot Spots

- Cancer genome anatomy project
- Clusters of orthologous groups
- Coffee Break
- Electroni...
- ...ou...
- Human map vie...
- Human/mouse homology maps
- LocusLink
- Malaria genetics & genomics
- Mouse genome resources
- ORF finder
- Reference sequence project